Computational Systems Biology **Fall Semester, 2007/8**

Final project

Determining protein function is one of the most challenging problems faced by today's computational biologists. The current knowledge on the functionality of proteins is fairly limited in most model organisms which makes the use of computational tools for completing of the missing infromation of primary importance. There are various methods in the literature for deducing protein function based on the current biological data. Such relevant data may include sequence similarity, protein-protein interactions (PPI), genetic interactions (e.g.. the so called "synthetic lethality"), correlation in expression patterns and more. The common theme in most methods regardless of the type of data being used is to rely on proteins whose function is already known in order to annotate those with an unknown function.

The goal of this exercise is to develop methods for assigning proteins to functional classes on the basis of their network of physical interactions (the *PPI network*). For a review on current methods for function prediction based on protein network information see [1].

In this project, you are given a PPI network and an annotation of a fraction of the proteins in the network. Your goal is to create the best possible predictor for the functionality of the remaining ones. You can use anything learnt throughout the course or reviewed in [1] and try as many different methods and different parameters as you wish. The more the better.

Regardless of which methods you choose, you should also implement one of the *direct methods* and one of the *graph theoretic* methods reviewed in [1].

The project should be submitted in groups of 2 or 3 no later than 1/7/08.

The dataset

The dataset you will be given consists of the PPI network and protein functional annotation of the yeast *Saccharomyces cerevisiae*.
The format of the files is as follows:

a. *net.txt* - The yeast PPI network with the following format:
[protein-name 1] [protein-name 2] [interaction confidence]

The first two columns are the names of the interacting proteins, the last column is a value between 0 and 1 denoting how confident we are that this interaction is real, with 1 meaning this interaction is indeed real (see [2] for how these values are determined).

b. *annot.txt* - protein annotation file of the following format:
[protein-name] [list of annotations]

The dataset will be available for download from the course web site at the beginning of next week.

Note that all input/ output files are (or should be) **tab-delimited**.

What to submit

1. A summary consisting of

a. A description of all the methods you used. You should implement at least two more methods in addition to the two compulsory ones. Note that the methods surveyed in [1] are only a suggestion. You may implement your own methods as long as you can explain why they should work well. You can also try to extend some of the surveyed methods, play with their parameters etc.

b. Assessment of accuracy of the different methods based on 10-fold cross validation. Here you should use *precision* and *recall* as your performance measures (see [1]).

c. Comparison between the different methods. Here you should explain what were the reasons for choosing the methods you implemented, and for the difference in performance between the methods. Specific examples (for specific proteins) highlighting conceptual differences between methods would be appreciated.

2. A decision of which method you think will best perform on the un annotated proteins (the *test set*), and its expected performance.

3. A file named *project.out*, consisting your predictions for the functions of the test set. This should be a tab-delimited text file of the same format as the *annot.txt* input file.

4. A linux executable called *fpredict* with the following interface:

./fpredict -n net -a annot -o outname

The input arguments are:

net - a network file, with the same format as *net.txt*.

annot - protein annotation file, with the same format as *annot.txt*.

out - the name of the output file to be produced by your program, with the same format as *annot.txt*. Note that your program should produce an output similar to *project.out* when provided with the input files *net.txt* and *annot.txt*.

5. All your source code, in a readable form with adequate indentation and complete documentation.

All the above should be submitted as a single *.tar* file.
You will also have to prepare a 15 minute presentation of your work. We will hold two sessions for these presentations in early July. Exact dates and places will be given in due time.

Grading

Grades will be given based on the amount of work invested in this project, clarity and insight of the summary, and the accuracy in prediction of the test set.

References

[1] Sharan R, Ulitsky I, and Shamir R. Network-based prediction of protein function. *Mol Sys. Biol.*, 3:88, 2007.

[2] Bader, J. S., Chaudhuri, A., Rothberg, J. M., & Chant, J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, **22**(1), 78–85.